Comments on "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging"

Thomas M. Hamill

NOAA Earth System Research Laboratory, Physical Sciences Division

Boulder, Colorado

Submitted to *Monthly Weather Review*

14 July 2006

Corresponding author address

Dr. Thomas M. Hamill
NOAA Earth System Research Laboratory
Physical Sciences Division
R/PSD 1, 325 Broadway
Boulder, CO 80305

1. **Introduction**.

Wilson et al. (2006, hereafter W06) recently described the application of the Bayesian model averaging (BMA, Raftery et al. 2005, hereafter R05) calibration technique to surface temperature forecasts using the Canadian ensemble prediction system. The BMA technique as applied in W06 produced an adjusted probabilistic forecast from an ensemble through a two-step procedure. The first step was the correction of biases in individual members. The second step was the fitting of a Gaussian kernel around each member of the ensemble. The amount of weight applied to each member's kernel and the width of the kernel(s) were typically set through an Estimation-Maximization (EM) algorithm (Dempster et al. 1997). The final probability density function (pdf) was a sum of the weighted kernels.

W06 reported (their Fig. 2) that at any given instant, a majority of the ensemble members were typically assigned zero weight, while a few select members received the majority of the weight. Which members received large weights varied from one day to the next. These results were counter-intuitive; why discard the information from so many ensemble members? Why should one member have value one day and none the next?

This comment to W06 will show that BMA based on EM is not an appropriate technique to use with small sample sizes; specifically, the radically unequal weights of W06 exemplify an overfitting (Wilks 2006a, p. 206) to the training data. In W06, the EM algorithm was required to set the weights of 16 individual ensemble members with 25-80 days of data.

To illustrate this point, a recent reforecast data set was used. This was comprised of many decades of daily ensemble forecasts with perturbed initial conditions and a

single forecast model. This large data set permitted a comparison of techniques with small and very large training samples. This reforecast data set used a T62, circa 1998 version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). A 15-member forecast, consisting of a control and 7 bred pairs (Toth and Kalnay 1997) was integrated to 15 days lead for every day from 1979 to current. For more details on this reforecast data set, please see Hamill et al. (2006). The verification data was the NCEP/NCAR reanalysis (Kalnay et al. 1996).

2. **Overfitting with the BMA-EM algorithm**.

EM is an iterative algorithm that adjusts the BMA model parameters (here, the weights, plus the standard deviation to apply to each kernel) through a two-step procedure of parameter estimation and maximization. R05 (eqs. 5-6, and accompanying text) provides details. The algorithm iterates to convergence, stopping when the change in log-likelihood function from one iteration to the next is less than a cutoff δ. The magnitude of δ may be chosen by the user, but is assumed $\delta << 1.0$.

To illustrate the tendency for the BMA EM to overfit when trained with small sample sizes, consider 4-day 850 hPa temperature ensemble forecasts for a grid point near Montreal, Canada. Forecasts were produced and validated for the 23 years * 365 days = 8395 cases. Because we would like to assume in this example *a priori* that the member weights should be equal, the 15-member ensemble was thinned, eliminating the slightly more accurate control member. The BMA algorithm was then trained using the remaining 14 identically distributed bred members and only the prior 40 days' forecasts and analyses, posited in W06 to be an acceptably long training period (note: here, the 40-

day training data for early in the first year of the data set used data from the end of year 23). In addition, the BMA algorithm was also trained in a cross validated manner using 22 years*91 days of data, with the 91 days centered on the Julian day of the forecast.

The BMA algorithm was coded generally following the algorithm used in the R05 and W06 articles. Two adjustments were used, however. First, no refinement to maximize the CRPS was performed, as suggested in R05. This had minimal impact. Second, W06's proposed regression correction was applied only to the ensemble mean, while the original deviation of each member about the mean was preserved. More concretely, given an ensemble member x_i^f, an ensemble mean $\bar{x}^f$, and a regression-corrected ensemble-mean forecast $\left(a + b\bar{x}^f\right)$, the member forecast was replaced with a forecast that was the sum of the initial perturbation from the ensemble mean and the corrected forecast:

$$x_i^f \leftarrow \left(x_i^f - \bar{x}^f\right) + \left(a + b\bar{x}^f\right) \quad . \tag{1}$$

This alternative regression approach was used because when every member was regressed separately, as forecast lead increased, all the members were regressed toward climatology and the ensemble spread of adjusted members shrank (see also Wilks 2006b). This is clearly an undesirable property; the spread should asymptotically approach the climatological spread of the ensemble forecast.

We now consider the properties of the EM algorithm. The initial guess for all member weights was 1/14. Keeping track of the ratio of maximum to minimum BMA member weights for each of the 8395 cases, these ratios were sorted, and the median ratio was plotted as the EM convergence criterion δ was varied. For the 40-day training

period, when $\delta = 0.01$, the largest and smallest weights were much more similar compared to when δ was tightened (Fig. 1a). With the 22 years of training data, the weights stayed much more equal as δ was tightened (Fig. 1b).

Could the unequal weightings with the small training set and tight δ actually be realistic? As mentioned in R05, as the EM iterates, the log-likelihood *of the fit to the training data* is guaranteed to increase. However, we can also track the fit to the validation data. Figures 2 a-b show the average training and validation log likelihoods (per forecast day) for the small and large training data sizes. Notice that for the small sample size, the validation data log likelihood gets smaller as the convergence criterion is tightened, a sign that the unequal weights are not realistic. The same effect is hardly noticed with the large training data set, where the weights remain nearly equal. This demonstrates that the highly variable weights are most likely an artifact of overfitting. Perhaps the assumption of independence of forecast errors in space and time (R05, p. 1159) is badly violated with these ensemble forecasts, and the EM algorithm actually is quite sensitive to this assumption. We agree with W06 proposition that the radical differences in weights are probably due to co-linearity of members' errors in the training data. What is clear here is that this co-linearity is largely is not properly estimated from small samples, leading to the inappropriate de-weighting and exclusion of information from some members.

Is there a way of setting the BMA weights to avoid radically de-weighting some members? If co-linearity of member errors in the training data were essentially zero, then the weights may resemble those that would be set in a weighted least-square process. Suppose the training data establishes that the estimated root-mean-square errors for the

bias-corrected members are $s_1, \ldots s_n$. The weights that would produce the minimum-variance estimate of the mean state (e.g., Daley, p. 36, eq. 2.2.3) under assumptions of normality of errors is

$$w_i = \frac{1}{s_i^2} \Big/ \sum_{j=1}^{n} \frac{1}{s_j^2} \quad . \tag{2}$$

The advantage of this method for setting weights, also, is that if there truly is a strong co-linearity of member errors, the BMA pdf should not be any worse as a consequence of using the more equal weights of eq. (2) rather than the unequal weights from a highly iterated EM. This can be demonstrated simply by considering two member highly co-linear forecasts with similar errors and biases, so $x_i^f \cong x_j^f$. Then the weighted sums are similar, regardless of the partitioning of the weights. For example,

$1.0 \times x_i^f + 0.0 \times x_j^f \cong 0.0 \times x_i^f + 1.0 \times x_j^f \cong 0.5 \times x_i^f + 0.5 \times x_j^f$.

4. **Conclusions.**

While the BMA technique is theoretically appealing, for ensemble forecast calibration, the EM technique cannot be expected to set realistic weights for each member when using a short training data set. Enforcing more similar weights among BMA members [eq. (2)] may work as well or better than the EM method.

References

Daley, R., 1986: *Atmospheric Data Analysis*. Cambridge Press, 457 pp.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc*., **39B**, 1-39.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc*., **87**, 33-46.

Kalnay, E., and coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc*., **77**, 437-472.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev*., **133**, 1155-1174.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*., **125**, 3297-3319.

Wilks, D. S., 2006a: *Statistical Methods in the Atmospheric Sciences* (2nd Edition). Academic Press, 627 pp.

---------- , 2006b: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Apps*., in press. Available from dsw5@cornell.edu.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2005: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging. *Mon. Wea. Rev*., accepted pending revisions. Available from lawrence.wilson@ec.gc.ca

FIGURE CAPTIONS

Figure 1: Log_{10} of the median sample's maximum member weight divided by minimum member weight, as a function of the EM convergence criterion. The median represents the (23*365/2)th rank-ordered ratio among the 23*365 sample days. (a) 40-day training period, (b) 22-year cross-validated training period.

Figure 2: Log likelihood (per unit day) of training and validation data as a function of the convergence criterion. (a) 40-day training data, and (b) 22-year cross-validated training data.

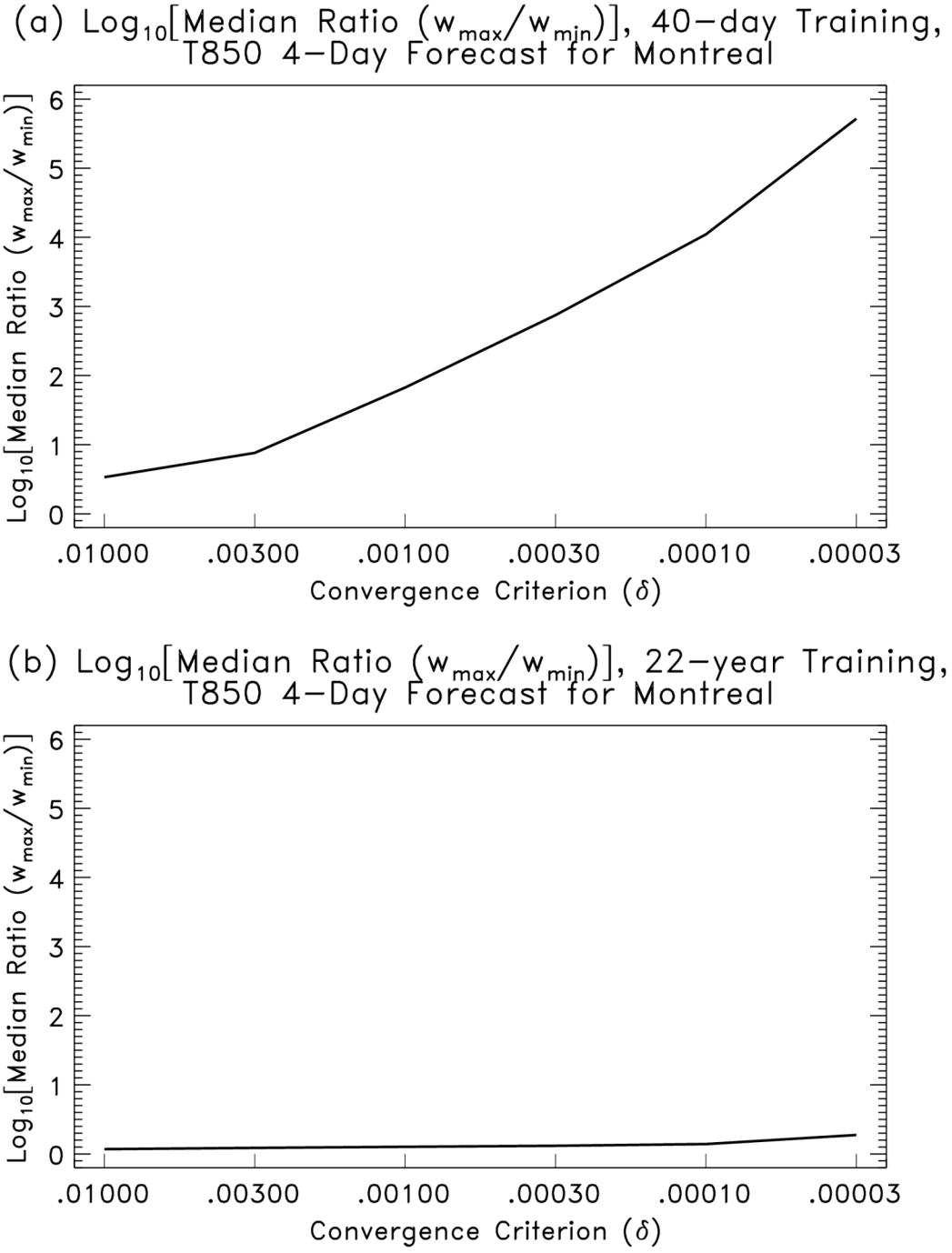

Figure 1: Log_{10} of the median sample's maximum member weight divided by minimum member weight, as a function of the EM convergence criterion. The median represents the (23*365/2)th rank-ordered ratio among the 23*365 sample days. (a) 40-day training period, (b) 22-year cross-validated training period.

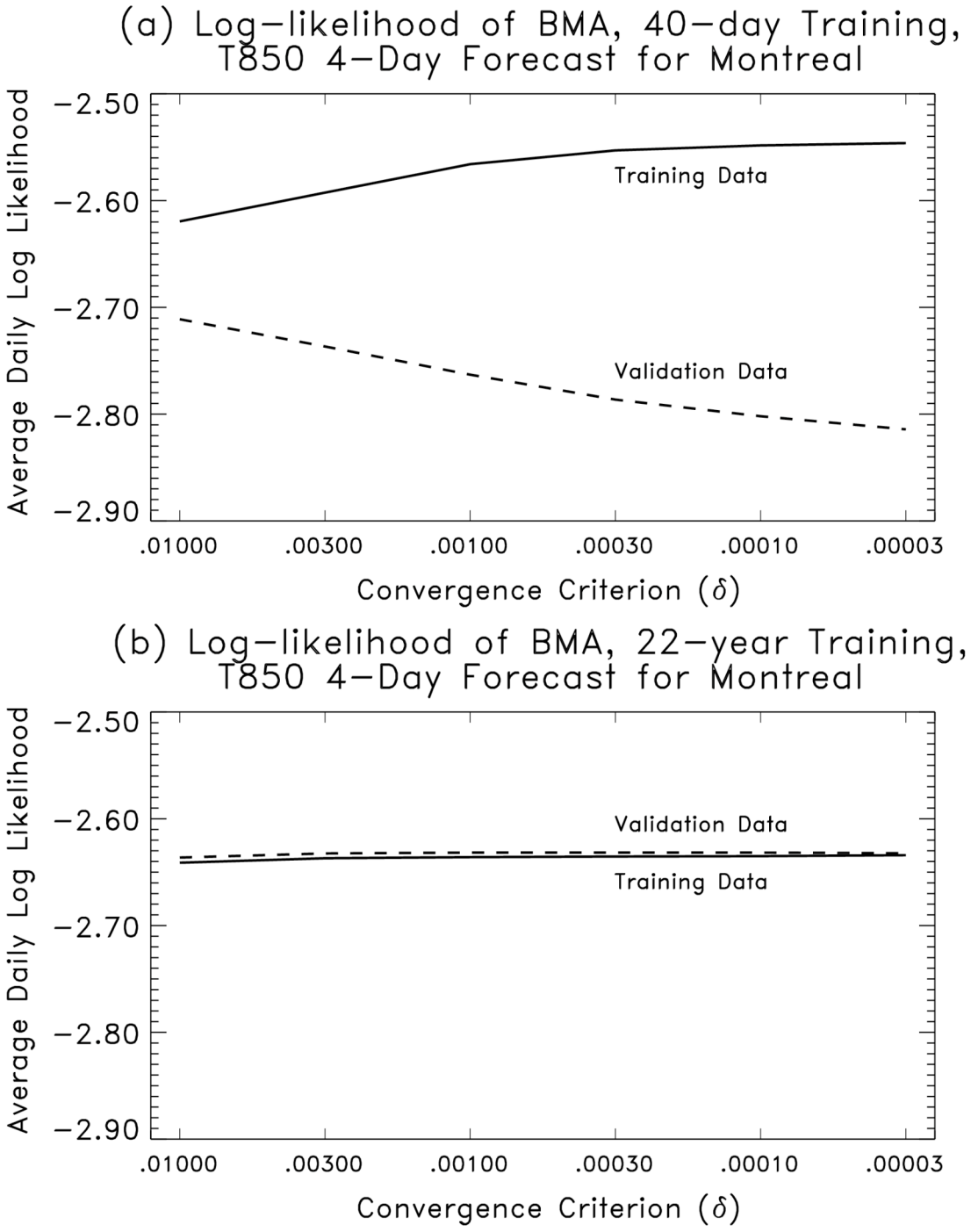

Figure 2: Log likelihood (per unit day) of training and validation data as a function of the convergence criterion. (a) 40-day training data, and (b) 22-year cross-validated training data.